

14 July 2016

# On a Certain Fallacy (probably committed by Gödel)

by

KAAVE LAJEVARDI

(La Société des Philosophes Chômeurs, Téhéran)

and

SAEED SALEHI

(University of Tabriz &amp; IPM)

**Abstract.** We take an argument of Gödel’s from his groundbreaking 1931 paper, generalize it, and examine its (in)validity. The argument in question is this: *A says of itself that it has property  $F$ , and  $A$  does have property  $F$ ; therefore,  $A$  is true.*

**§1. Gödel’s Introduction.** As is well known, Gödel begins his 1931 masterpiece with an introductory section, Section 1, wherein he explains the main ideas behind the (first) incompleteness theorem in an informal and intuitive way, with an explicit caveat that those remarks are being made “without any claim to complete precision”. What he does in that section is, among other things, to introduce, in a very lucid way, the idea of encoding the syntax and that of diagonalization.

But Gödel’s introduction is not confined to what can be found, in a more formal expression, in the technical parts of the paper. Gödel’s official statement of the first incompleteness theorem (his Theorem VI) asserts that for every theory satisfying certain conditions, there is a sentence, now called the Gödel sentence of the theory, which is expressible in the language of the theory but is neither provable nor refutable in the theory—the theorem does *not* talk about the truth of that undecidable sentence. In his Section 1, however, Gödel chooses to also talk about the truth of the Gödel sentence of the system of *Principia Mathematica*, and observes that the proof of the first incompleteness theorem, as presented in that section, is similar to Richard’s paradox. Whatever expository merits it might have, this move is a problematic one, and Gödel shows awareness of this.<sup>1</sup>

Our purpose in this note is not to philosophize about the notion of truth or doing exegetical work on its rôle in Gödel’s classic paper. Rather, we wish to draw attentions to an informal argument, to the effect that the Gödel sentence of the system *Principia Mathematica* is true (that is, true in  $\mathbb{N}$ ), which is presented by Gödel in his introductory section. We dare to claim that the argument is invalid.

---

<sup>1</sup> Gödel says that his method of proof is applicable to any formal system that has two conditions, the second of which being “every provable formula is *true* in the interpretation considered” (p. 151, emphasis added). He then writes, “The purpose of carrying out the above proof with full precision in what follows is, among other things, *to replace the second of the assumptions just mentioned by a purely formal and much weaker one*” (ibid., emphasis added).

As it goes without saying, this will do no harm to the veracity of Gödel’s celebrated theorems; it just shows that the greatest minds can make mistakes, especially when it comes to informal discussions with a “philosophical” flavor. Even if we fail to convince our readers that Gödel’s argument is in fact invalid, we hope at least help them to decipher what the Master had in mind when he wrote it.

The argument in question goes as follows (K. GÖDEL, 1931, p. 151) (*italics in the original*):

From the remark that  $[R(q); q]$  says about itself that  
it is not provable, it follows at once that  $[R(q); q]$  is true,  
for  $[R(q); q]$  *is* indeed unprovable (being undecidable).

Here  $[R(q); q]$  is the **Gödel sentence**<sup>2</sup> of a theory (or “system”) which is subject to the first incompleteness theorem. From now on, let us call it **G**.

As we understand the above passage, its logical form is the following, where  $A$  is an arbitrary sentence expressible in the language of the theory:

- (1)  $A$  says about itself that it has a property  $F$ .
- (2)  $A$  indeed has the property  $F$ .
- (3) Therefore:  $A$  is true.

That Gödel says that it follows *at once*<sup>3</sup> that **G** is true suggests that, in Gödel’s view, we are not dealing with an enthymeme—it seems to us that, for Gödel, the argument scheme has no missing premises. It is our task in the next section to argue that the (1)–(3) argument scheme is **invalid**, that is to say, there are situations wherein the premises are true while the conclusion false.

Any invalid argument could be rendered valid by the addition of some premises. So, in order to strengthen our case, and in order to be maximally fair to Gödel, we shall understand the above argument augmented with other premises. Most importantly, the context of Gödel’s introduction almost forces us to presume that the system is consistent.<sup>4</sup> With an eye to his formulation of the first incompleteness theorem, one might also feel an obligation to presume something even stronger than simple consistency: we assume that the system is  $\omega$ -consistent.

In the next section we proceed to support our claim about the invalidity of the (1)–(3) argument. We assume that we are working with an  $\omega$ -consistent recursively axiomatizable extension of Robinson’s **Q**. (The system *PM* of Gödel’s title is of course axiomatizable; hopefully it is  $\omega$ -consistent as well.) Naturally enough, the validity of the argument hinges on its terms, in particular on what it is for a

<sup>2</sup> i.e., a sentence  $P$  which is provably equivalent to  $\neg\text{Pr}(\#P)$ , where  $\text{Pr}$  is the provability predicate of the theory and  $\#P$  is the standard term for the Gödel number of  $P$ . As all such  $P$ s are provably equivalent under certain minimal conditions normally met by the theories under consideration, we are allowed to talk about *the* Gödel sentence of a given theory.

<sup>3</sup> Gödel’s term (“sofort”) could also be translated to *immediately*.

<sup>4</sup> If the system is inconsistent, then **G** is actually false (in the standard model), but demonstrating this fact is not as easy as it might appear. Recall that the usual way of arguing for the truth of **G** is to use the fact that **G** was constructed, via the diagonal lemma, in a way that the system proves  $\mathbf{G} \leftrightarrow \neg\text{Pr}(\#\mathbf{G})$ . However, under the assumption of inconsistency, this equivalence is totally useless in determining the truth-value of **G** (in  $\mathbb{N}$ , or anywhere), for the system proves any other thing as well, e.g.,  $\mathbf{G} \leftrightarrow \text{Pr}(\#\mathbf{G})$ . One has to look at the details of the construction of **G** in order to determine the truth-value (in  $\mathbb{N}$ ) of the Gödel sentence of an inconsistent theory.

sentence to be “true”, and what is meant by a sentence “saying of itself” that it has a certain property. As for the first term, it is almost obvious from Gödel’s introductory section that, like many modern writers in mathematical logic, when he writes “true” simpliciter, he means *true in the standard model*,  $\mathbb{N}$ .<sup>5</sup> Regarding the notion of saying something of oneself, we shall consider two interpretations that might be ascribed to Gödel.

**§2. The Invalidity.** How are we to understand the expression “ $A$  says that it has the property  $F$ ”? Modulo an agreement over the meaning of “holding”, to which we shall return shortly, we think it is uncontroversial that if the conditional  $A \rightarrow F(\#A)$  holds, then  $A$  says, *inter alia*, that it has property  $F$ . Thus to say that  $A$  says, *exactly*, that it has property  $F$  is to say that the biconditional  $A \leftrightarrow F(\#A)$  holds. And this is, in fact, what is taken to be the meaning of “ $A$  says of  $A$  that it has property  $F$ ” in the literature; see, e.g., P. MILNE (2007). As for the meaning of “holding”, we already presented evidence for the claim that what Gödel meant by it is being true in the standard model,  $\mathbb{N}$ . However, let us recognize another reading of it—which is also suggested in the literature (see P. MILNE (2007); V. HALBACH & A. VISSER (2014) and references therein)—namely *being provable in a given theory*. Having fixed a theory (which we suppose to be an  $\omega$ -consistent recursively axiomatizable extension of  $\mathbf{Q}$ ), we then have eight possible ways of interpreting the (1)–(3) argument scheme. Here is the complete list, of which we find (VII), (V), and (III) the most interesting:

$$\begin{array}{ll}
 \text{(I)} \frac{\mathbb{N} \models A \leftrightarrow F(\#A), \quad \mathbb{N} \models F(\#A)}{\mathbb{N} \models A} & \text{(II)} \frac{\mathbb{N} \models A \leftrightarrow F(\#A), \quad \mathbb{N} \models F(\#A)}{T \vdash A} \\
 \text{(III)} \frac{\mathbb{N} \models A \leftrightarrow F(\#A), \quad T \vdash F(\#A)}{\mathbb{N} \models A} & \text{(IV)} \frac{\mathbb{N} \models A \leftrightarrow F(\#A), \quad T \vdash F(\#A)}{T \vdash A} \\
 \text{(V)} \frac{T \vdash A \leftrightarrow F(\#A), \quad \mathbb{N} \models F(\#A)}{\mathbb{N} \models A} & \text{(VI)} \frac{T \vdash A \leftrightarrow F(\#A), \quad \mathbb{N} \models F(\#A)}{T \vdash A} \\
 \text{(VII)} \frac{T \vdash A \leftrightarrow F(\#A), \quad T \vdash F(\#A)}{\mathbb{N} \models A} & \text{(VIII)} \frac{T \vdash A \leftrightarrow F(\#A), \quad T \vdash F(\#A)}{T \vdash A}
 \end{array}$$

Of these, (I) and (VIII) are of course valid because of the truth-condition of the material conditional and Modus Ponens, respectively. For all other cases, we will present triples  $(A, F, T)$  which invalidate them.

Recall that “ $\mathbf{Q}$ ” denotes Robinson’s Arithmetic, and “ $\mathbf{G}$ ” denotes its Gödel sentence, i.e., we have  $\mathbf{Q} \vdash \mathbf{G} \leftrightarrow \neg \text{Pr}_{\mathbf{Q}}(\# \mathbf{G})$  and also  $\mathbb{N} \models \mathbf{G} \leftrightarrow \neg \text{Pr}_{\mathbf{Q}}(\# \mathbf{G})$ , where  $\text{Pr}_{\mathbf{Q}}(x)$  is the provability predicate of  $\mathbf{Q}$ .<sup>6</sup>

<sup>5</sup> Thus K. GÖDEL (1931)’s footnote 4 on page 145: “... no other notions occur but  $+$  (addition) and  $\cdot$  (multiplication), both for natural numbers, and in which the quantifiers  $(x)$ , too, apply to natural numbers only.”

<sup>6</sup> It is perhaps worth mentioning that here we do not assume that  $\mathbf{Q}$  is sound. For example, by the construction of  $\mathbf{G}$  one can rather easily show (i)  $\mathbb{N} \models \mathbf{G} \leftrightarrow \neg \text{Pr}_{\mathbf{Q}}(\# \mathbf{G})$ , and with a longer (and more delicate) proof one can also derive (ii)  $\mathbf{Q} \vdash \mathbf{G} \leftrightarrow \neg \text{Pr}_{\mathbf{Q}}(\# \mathbf{G})$ ; so we do not infer (i) from (ii) *and the soundness of  $\mathbf{Q}$* .

**THEOREM 2.1.** *The argument (IV) is invalid for  $A = \mathbf{G}$ ,  $F(x) \equiv (x = \# \mathbf{G})$ , and  $T = \mathbf{Q}$ .*

*Proof.* Obviously,  $\mathbb{N} \models F(\# \mathbf{G})$ . Also, because of the expressive power of  $\mathbf{Q}$ , we have  $\mathbf{Q} \vdash F(\# \mathbf{G})$ . By Gödel's proof  $\mathbb{N} \models \mathbf{G}$  holds and so  $\mathbb{N} \models \mathbf{G} \leftrightarrow F(\# \mathbf{G})$ . On the other hand, by Gödel's theorem,  $\mathbf{Q} \not\vdash \mathbf{G}$ .  $\square$

**THEOREM 2.2.** *The following show that (II) and (VI) are invalid:  $A = \mathbf{G}$ ,  $F(x) \equiv \neg \text{Pr}_{\mathbf{Q}}(x)$ , and  $T = \mathbf{Q}$ .*

*Proof.* We already have  $\mathbb{N} \models \mathbf{G} \leftrightarrow \neg \text{Pr}_{\mathbf{Q}}(\# \mathbf{G})$ . Since by Gödel's theorem we have  $\mathbf{Q} \not\vdash \mathbf{G}$ , it follows that  $\neg \text{Pr}_{\mathbf{Q}}(\# \mathbf{G})$  is true, i.e.,  $\mathbb{N} \models \neg \text{Pr}_{\mathbf{Q}}(\# \mathbf{G})$ .  $\square$

Let  $\omega\text{-Con}_{\mathbf{Q}}(x)$  be a formula of the language of arithmetic which expresses the following: *if the formula with Gödel number  $x$  is appended to  $\mathbf{Q}$ , the resulting theory is  $\omega$ -consistent* (see (D. ISAACSON, 2011, Proposition 19) for the details). Let  $\mathbf{K}$  be the fixed-point of  $\neg \omega\text{-Con}_{\mathbf{Q}}(x)$  obtained by the diagonal lemma—that is to say, let  $\mathbf{K}$  be such that  $\mathbf{Q} \vdash \mathbf{K} \leftrightarrow \neg \omega\text{-Con}_{\mathbf{Q}}(\# \mathbf{K})$  and also  $\mathbb{N} \models \mathbf{K} \leftrightarrow \neg \omega\text{-Con}_{\mathbf{Q}}(\# \mathbf{K})$ . As shown by ISAACSON (ibid., with a proof attributed to Kreisel in the 1950s), the  $\Sigma_3$ -sentence  $\mathbf{K}$  is false (in  $\mathbb{N}$ ) and the theory  $\mathbf{Q} + \mathbf{K}$  is  $\omega$ -consistent.

**THEOREM 2.3.** *The argument (V) does not hold for  $A = \mathbf{K}$ ,  $F(x) \equiv (x = \# \mathbf{K})$ , and  $T = \mathbf{Q} + \mathbf{K}$ .*

*Proof.* By  $\mathbf{Q} \vdash F(\# \mathbf{K})$  we have  $\mathbf{Q} + \mathbf{K} \vdash \mathbf{K} \leftrightarrow F(\# \mathbf{K})$ . Now,  $\mathbb{N} \models F(\# \mathbf{K})$  holds trivially. By (D. ISAACSON, 2011, Propositions 19),  $\mathbb{N} \not\models \mathbf{K}$ .  $\square$

**THEOREM 2.4.** *Neither (III) nor (VII) hold for  $A = \mathbf{K}$ ,  $F(x) \equiv \neg \omega\text{-Con}_{\mathbf{Q}}(x)$ , and  $T = \mathbf{Q} + \mathbf{K}$ .*

*Proof.* We already have  $\mathbf{Q} \vdash \mathbf{K} \leftrightarrow \neg \omega\text{-Con}_{\mathbf{Q}}(\# \mathbf{K})$  and  $\mathbb{N} \models \mathbf{K} \leftrightarrow \neg \omega\text{-Con}_{\mathbf{Q}}(\# \mathbf{K})$  by definition, and so  $\mathbf{Q} + \mathbf{K} \vdash \mathbf{K} \leftrightarrow \neg \omega\text{-Con}_{\mathbf{Q}}(\# \mathbf{K})$  holds too. The latter also implies that  $\mathbf{Q} + \mathbf{K} \vdash \neg \omega\text{-Con}_{\mathbf{Q}}(\# \mathbf{K})$ . Finally, by (D. ISAACSON, 2011, Propositions 19) we have  $\mathbb{N} \not\models \mathbf{K}$ .  $\square$

Out of fairness to Gödel, let us acknowledge the fact that perhaps it is only by overgeneralizing his informal argument that we are making the argument invalid. Had we not abstracted from the specific properties of  $A$ ,  $F$ , and  $T$ , Gödel's informal argument would be valid, even in the interesting cases of (III), (V), and (VII) (though it would lose much of its appeal). This is substantiated in the following:

**PROPOSITION 2.5.** *If  $A, F \in \Pi_1$  and the  $\omega$ -consistent theory  $T$  is  $\Sigma_1$ -complete, then the arguments (III), (V), and (VII) are valid.*

*Proof.* Let  $A = \forall x \theta(x)$  for some  $\theta \in \Sigma_0$  and  $\mathbb{N} \not\models A$ . Then there exists an  $m \in \mathbb{N}$  such that  $\mathbb{N} \models \neg \theta(\bar{m})$ . So,  $T \vdash \neg \theta(\bar{m})$ , thus  $T \vdash \neg A$ . Now we reach a contradiction in each case:

- (III): from  $\mathbb{N} \models \neg F(\# A)$  and  $\neg F(\# A) \in \Sigma_1$  we have  $T \vdash \neg F(\# A)$ , which contradicts the assumption  $T \vdash F(\# A)$ .
- (V): from  $T \vdash \neg A$  we have  $T \vdash \neg F(\# A)$  and then, since  $\neg F \in \Sigma_1$  and  $T$  is  $\omega$ -consistent, we should have  $\mathbb{N} \models \neg F(\# A)$ ; contradiction.
- (VII): similar to the case of (V), we have  $T \vdash \neg F(\# A)$ , contradicting  $T \vdash F(\# A)$  in the assumptions.  $\square$

Let us note that if  $T$  is a sound theory then (III), (V), and (VII) are of course valid; Proposition 2.5. shows that we can replace the condition of soundness of  $T$  with its  $\omega$ -consistency and  $\Sigma_1$ -completeness.

**§3. Concluding Remarks.** Where does this leave us? It is a fact that, in the final paragraph of his introductory section, Gödel gives an argument, which we quoted in full. Is that argument valid? We took (1)–(3) to be the form of the argument, and we envisaged eight ways to understand it, of which we showed six to be invalid.

1. Of the two valid arguments, we find (I) to be rather insipid and not in accord with the flavor of Gödel’s overly syntactic approach in the technical part of his paper. However, we are not going to argue for this here. If Gödel’s intended reading of the argument was (I), let it be.

As for (VIII), which is the only other valid one, it is an objective fact that no ideally intelligent logician could take it to be what the (1)–(3) argument means. This is because, we recall, Gödel’s property  $F$  is that of unprovability in the system and he says that  $\mathbf{G}$  does have the property  $F$ . Now if we think that by (2) he meant that  $F(\#A)$  is provable in  $T$ , we should attribute the idea to Gödel that the system proves  $\neg\text{Pr}(\#\mathbf{G})$ . But it is an easy consequence of Löb’s theorem that no consistent theory proves the unprovability of *anything*.<sup>7</sup> So that, at least with the wisdom of hindsight, Gödel had better not thought of “ $\mathbf{G}$  is indeed unprovable” to mean that the system proves the unprovability of  $\mathbf{G}$  (though it is not clear to us if Gödel was aware of Löb’s result). This observation rules out the reading (VIII) (and of course those of (III), (IV), and (VII) as well). Perhaps we should be content that we have made it clear what Gödel meant by the argument.

2. As we noted after Proposition 2.5., the interesting cases of (III), (V), and (VII) turn out to be valid if  $T$  is sound. The assumption of soundness may appear to be what Gödel had in mind, for, in the very first paragraph of his paper, he says that his result holds in particular for every extension of  $PM$  and  $ZF$ , “provided no false proposition of the kind specified in footnote 4 become provable owing to the added axioms” (K. GÖDEL, 1931, pp. 145 f).<sup>8</sup> In itself, this strongly suggests that he is talking about **sound** theories, i.e., theories  $T$  with  $\mathbb{N} \models T$ . Also, in the final paragraph of his Section 1 he says—and we have quoted this in our footnote 1—that his method of proof is applicable to any formal system which is, first, of sufficient expressive power and, secondly, is such that “every provable formula is true in the interpretation considered”. This, too, may suggest (though not as forcibly as the previous quote) that Gödel is talking about *sound* theories. However, soundness cannot be what he had in mind as a required property for theories under consideration, for he immediately adds that the purpose of the technical part of the paper is, among other things, to replace this second condition by a purely formal

<sup>7</sup> If  $T$  proves  $\neg\text{Pr}(\#B)$ , then  $T$  proves  $\text{Pr}(\#B) \rightarrow B$  as well. Hence, by Löb’s theorem,  $T$  proves  $B$ . It follows that  $T$  proves  $\text{Pr}(\#B)$  which, together with our starting assumption, makes  $T$  inconsistent.

<sup>8</sup> Gödel’s footnote 4, which we partly quoted in our footnote 5, says, among other things, that “the quantifiers... apply to natural numbers only”.

and much weaker one (see our footnote 1 again). It is quite obvious that by the “much weaker” assumption he means  $\omega$ -consistency, a condition he introduces just before his Theorem VI (on page 173). Which brings us to our next remark.

3. We had already noted (Proposition 2.5.) that with certain extra premises concerning the complexity of  $A$  and  $F$  and the  $\omega$ -consistency and  $\Sigma_1$ -completeness of the system, what we find the most interesting readings of Gödel’s (1)–(3) argument are in fact valid. Yet, we find this a matter of reading too much into Gödel’s argument in that introductory section—recall that he says that the conclusion follows *at once*. At any rate, we consider it an open possibility that Gödel really meant to talk about  $\Pi_1$ -formulas and  $\Pi_1$ -properties and  $\omega$ -consistent,  $\Sigma_1$ -complete theories. If that is what he meant, let it be.

4. Perhaps more importantly, we are aware that there is an ongoing discussion on what it means for a formula to say something of itself (in this regard, the paper of V. HALBACH & A. VISSER (2014) is an invaluable source). By no means do we want to neglect the debate. However, if the task is to *evaluate Gödel’s argument as presented in his introductory section*, we take it obvious that he had thought of the very sentence  $\mathbf{G}$  (or  $[R(q); q]$  is his own notation) as a sentence saying of itself that it was unprovable. So, whatever the correct analysis of the concept of self-attribution might turn out to be, *what Gödel had in mind* must have been something retrievable from what he has done in his 1931 paper. Given that his apparatus for constructing  $\mathbf{G}$  is the diagonal lemma (K. GÖDEL, 1931, the formula  $p$  on page 175 in the proof of Theorem VI), we think there is no choice but to think that, for Gödel, saying that  $A$  says of itself that it has property  $F$  means either that  $A \leftrightarrow F(\#A)$  is provable in the system, or else this biconditional is true in  $\mathbb{N}$ . We therefore find (I)–(VIII) as the only readings of the (1)–(3) argument that Gödel could have had in mind.

#### Bibliography

- K. GÖDEL (1931). On Formally Undecidable Propositions of *Principia Mathematica* and Related Systems. I. in: *Kurt Gödel Collected Works, Volume I: Publications 1929–1936*, S. Feferman et al., eds. (Oxford University Press 1986) pp. 145–195.
- V. HALBACH & A. VISSER (2014). Self-reference in Arithmetic. I: *The Review of Symbolic Logic* 7(4) 671–691; II: *ibid* 692–712.
- D. ISAACSON (2011). Necessary and Sufficient Conditions for Undecidability of the Gödel Sentence and its Truth. in: *Logic, Mathematics, Philosophy, Vintage Enthusiasms: Essays in Honour of John L. Bell*, D. DeVidi et al., eds. (Springer 2011) pp. 135–152.
- P. MILNE (2007). On Gödel Sentences and What They Say. *Philosophia Mathematica* 15(2) 193–226.

(KAAVE LAJEVARDI)

LA SOCIÉTÉ DES PHILOSOPHES CHÔMEURS,

TEACHERS CAFÉ, 1337, RUE D’E’JAAZI, 198888 TÉHÉRAN, IRAN

E-mail: kaave.lajevardi@gmail.com

(SAEED SALEHI)

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF TABRIZ,

29 BAHMAN BOULEVARD, P.O.BOX 51666–17766, TABRIZ, IRAN

E-mail: salehipour@tabrizu.ac.ir, saeedsalehi@ipm.ir